Data Design and Representation

Final Project Report

By:

Vamsee Krishna Narahari,

Vindhya Mandekar,

Aishwarya

Table of Contents

Executive Summary 2
Background 3
Introduction to Data 4
Business Case 7
Conclusion 8
Appendix 9

Executive Summary

This project aims at creating a database of tourist activities, restaurants, and hotels in San Francisco. The websites Travel Advisor and Hotels.com have been scraped to collect the details of the top 300 destinations, 3801 restaurants and 500 hotels in San Francisco. In the end, it would be leveraged to find the nearest restaurant based on your current destination (vice-versa) and distance to your hotel.

The programming language chosen for this project is Python. Some of the libraries used are Beautiful soup, requests, selenium. The position stack api is used to fetch the latitude and longitude details of all data points. The database design of choice is Mongodb due to its flexible and scalable architecture. The database contains details such as name, address, star ratings etc. of things to do in San Francisco. For restaurants, the details captured are restaurant name, address, geo-location, and contact details. The name, ratings, amenities, geo-location and places-nearby are captured for hotels.

The thorough details saved in the database can be used by tourist and travelers to plan their itinerary in a seamless and optimized fashion. Visitors can plan to visit all nearby places at once and they can get this information by looking at the dashboard populated with details of all the tourist places using the database. The database collections can also be used by businesses to plan the location of their next venture. The areas which lack restaurants and hotels can be filtered out by looking at the visualization of all points from data collection and can be further analyzed.

Background

Preparing an itinerary for a vacation to any city requires a lot of effort. Starting from researching places to visit to creating a list of places to be covered in the same vicinity, it’s imperative to optimize the vacation time to get the most out of it. This project is aimed at creating a collection of data points that can be used by tour planners and travelers to plan their vacation better. The target industry is tourism and hospitality. Currently, data for tourist attractions, hotels and restaurants of San Francisco are collected in a database which can be expanded to include more cities. The details of various things to do in San Francisco are collected along with the available hotels and restaurants. All this information can be used by a tourist/traveler who is visiting a place and is looking for a hotel to stay or a restaurant to eat. This data can be highlighted on a map to help visualize the nearby places through map points. With all tourist places highlighted on a map, it is easier for a user to decide and cover nearby places in one day or plan their stay closer to the places they want to visit. The same applies to restaurants and hotels too. Instead of searching for a nearby restaurant or hotel, the user can look at all the details in a map visualization.

If planning to start a new business venture in the tourism industry, a lot of research and data collection is needed to square down on an area. It often takes a lot of analysis on existing businesses and their market grasp before pinpointing a location. Since, the database created contains details and geographic locations of restaurants and hotels, it can be used to create a dashboard and filter out the areas of interest. It will help save a preliminary analysis time and more effort can be put in to strategize the differentiation, location, market leadership and profit strategy.

Introduction to Data

The data is collected through Trip Advisor which is a leading travel company as well as Hotels.com. All information about things to do, hotel and restaurants in San Francisco are scraped from these websites.

For **things to do in SF**, a total of 300 tourist attractions are collected. The collected information contains name of the site, website address, star ratings, trip advisor ratings, top 3 customer reviews, image URL, description, admission cost, address and geolocation.

For **restaurants**, a total of 3801 restaurants are collected. The collected information contains the restaurant name and their website url, rating and number of reviews, cuisine and costs, address and geo-locations.

For **hotels**, a total of 500 hotels in and around San Francisco have been scraped from Hotels.com. The information collected include hotel name, address, star ratings, number of reviews, amenities, places nearby and the geolocation.

Web Scraping Routine:

The programming language used for web scraping is python. Some of the libraries used are Beautiful Soup, requests, json and selenium. The position stack api is used for collecting latitude and longitude from the address.

Steps of scraping for tourist attractions:

- A save_pages function is used to save the 10 pages for trip advisor tourist attraction to disk
- These saved pages are scraped to gather destination information along with their individual page urls using tourist_attractions function.

- Now the individual pages are saved to disk using individual_pages function. These pages are then accessed to scrape further details of each of these destinations using access_each_page function.
- The top 3 customer reviews from all these locations are also fetched and saved to to the database.
- A function names get_location is used to get geolocation of the destinations. These are indexed to be further leveraged for visualizing on a map.

Steps of scraping for restaurant names:

Selenium has been used since some of the data requires clicks on the buttons.

- A search function to launch and load the city and the content (restaurant or hotel etc) page to be loaded.
- Restaurants extract function to click on each page button and download that particular page, while capturing the URL and Name of the restaurant on the website
- Extract restaurant details function to extract each restaurant specific details from the URL in the second process.
- Post fetching the data, some of the text cleaning using regex to identify cost of the dollars, to pick specific details in the cuisine element etc
- In the extract Geo-Location function, the address of the restaurant has been requested to the position stack API to fetch the details of the Latitude and Longitude.
- Finally, the geo-locations have been appended into the Mongo dB in the format of the point, so that it would be easy to do the geo-location calculations. An index has been created on the geo-location as well.

Steps for scraping hotel names:

- A search function - get_search_results, launches the hotels.com website which has city chosen. It uses Selenium to click through and load all the results for that city.
- We then navigate through the results page to save all the urls for each hotel and then save the individual hotel pages to disk
- The extract_hotel_details function is used to extract all the necessary information from the page like hotel name, rating, address, amenities, and places nearby.
- We then use the position stack api to get the geo coordinates for each hotel's address.
- All of this information is stored in MongoDB in a collection called SFHotels. We have an index on the geolocation field.

Database design choices:

The database choice for this project is MongoDB due to the flexible nature of row length. Since, the review for each activity is being stored in the database, the size of each review may differ. There are also some entries where there is no review. We have decided to store the different datasets in separate collections as we would want to offer selection of only restaurants or hotels or tourist attractions irrespective of the others. Details of database and collection name are below:

Data	Database Name	Collection Name
Tourist attraction	Travelogue	sf_tourist_attraction
Restaurants	SanFranciscoTravelogue	final_trip_adv_rest
Hotels	SanFranciscoTravelogue	SFHotels

Business Case

The dataset provides a holistic view of the information needed for a traveler to make easier and optimized choices. The business can provide this information in the form of a map that contains all the details in a graphical format. A visual representation is easier for comprehension as well as promotes quick decision making. In another use-case, this data can be used by businesses to find a location that attracts most tourists. Areas that have tourist destinations and not enough restaurants or hotels would be a nice spot to start a new eatery. This visualization would help filter out a small area out of a big city to start a business. Further research can be done by businesses to identify a fitting place.

Similar platforms present today contain such information in the form of text. User must search for a specific location to see some results. There is no platform available today which shows all information in one viewport based on location proximity. Using this dataset, an interactive dashboard can be created with images of the place so user can make a choice quickly without getting into the specific pages of the location.

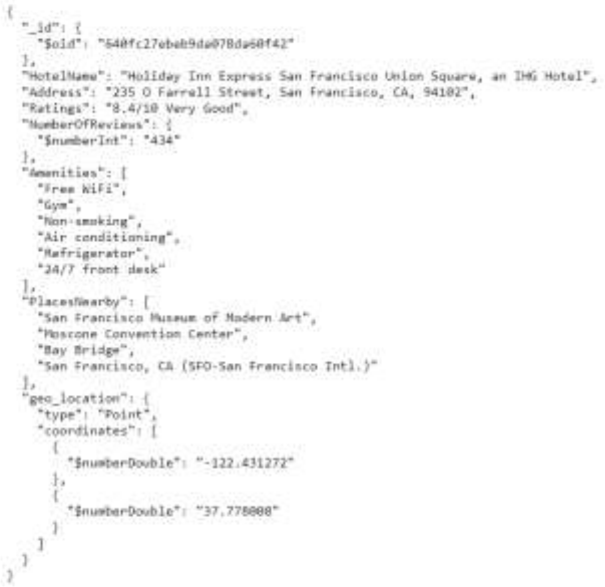
The choice of Mongodb is best because it adds allows to account for missing details for some of the locations. For example, not every tourist location in San Francisco has an admission fee. By using Mongodb, the admission can be added for those documents that have this entry and can be skipped for others. In cases of updates needed to the already existing locations, it can be done by directly accessing a location by its key and without changing the collection structure. The existing dataset can be expanded by adding details of more locations in the same city or from different cities without modifying the collection structure. If SQL was chosen, any modification would need a new column to be added. In such case, even the attributes which only apply to a few entries would take up memory for all the rows. Expanding an existing SQL table to account for additional tourist attractions from different cities may need us to keep adding more attributes, ultimately leading to too many attributes with fewer values in them.

Conclusion

In conclusion, the project creates a database encapsulating details of tourist attractions, restaurants, and hotels in San Francisco in a format which is easy to modify, append and scale to multiple cities. This information can be used to create a visual representation of all things in proximity helping tourists in their decision making. It can also be used to create other dashboards that can help businesses find their next target spot which attracts most travelers.

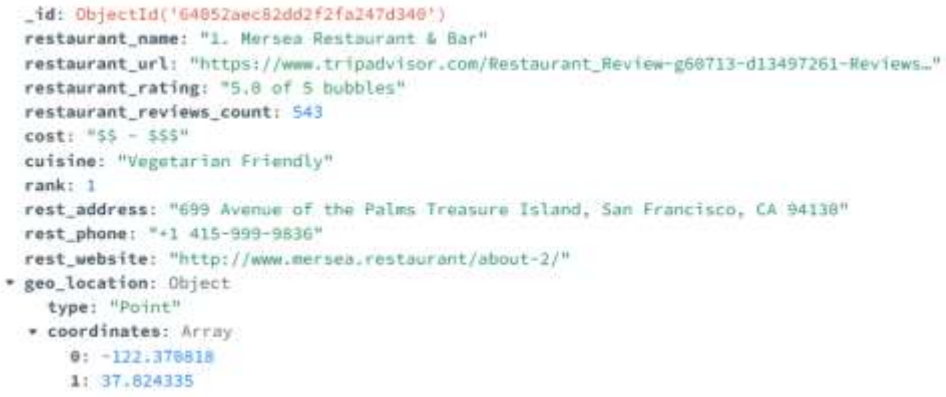
Appendix

Samples Json – Hotels



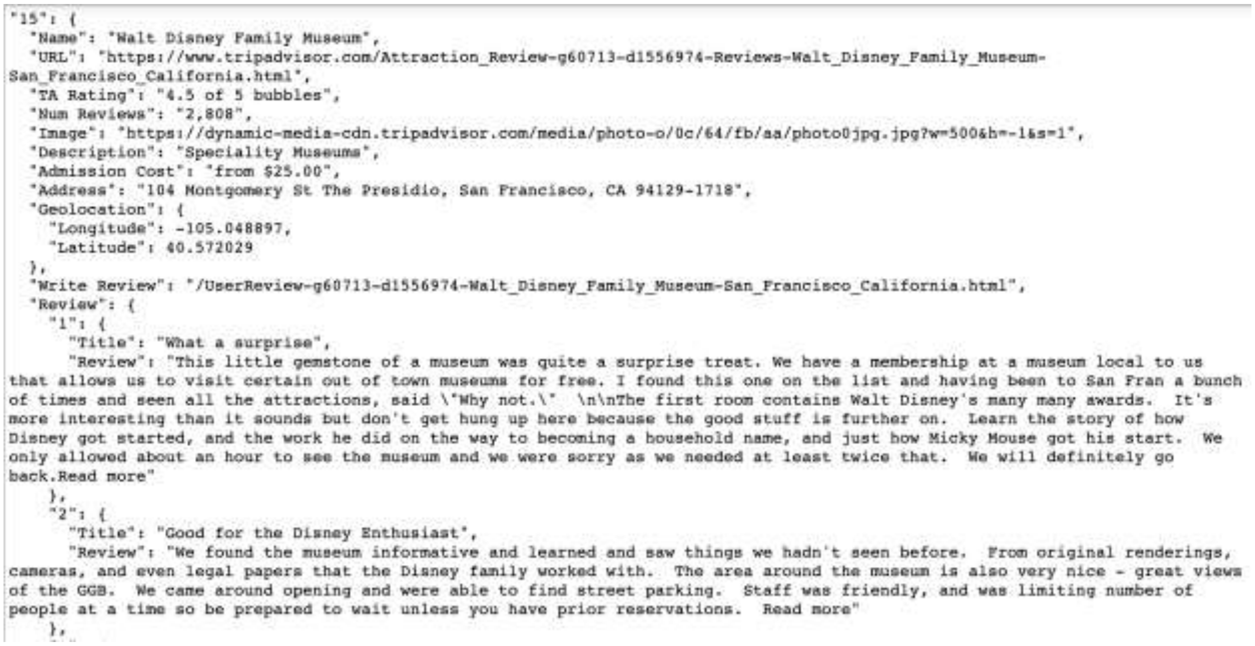
```
{
  "_id": {
    "$oid": "640fc27ebeb9da078da60f42"
  },
  "HotelName": "Holiday Inn Express San Francisco Union Square, an IHG Hotel",
  "Address": "235 O Farrell Street, San Francisco, CA, 94102",
  "Ratings": "8.4/10 Very Good",
  "NumberOfReviews": {
    "$numberInt": "434"
  },
  "Amenities": [
    "Free WiFi",
    "Gym",
    "Non-smoking",
    "Air conditioning",
    "Refrigerator",
    "24/7 front desk"
  ],
  "PlacesNearby": [
    "San Francisco Museum of Modern Art",
    "Moscone Convention Center",
    "Bay Bridge",
    "San Francisco, CA (SFO-San Francisco Intl.)"
  ],
  "geo_location": {
    "type": "Point",
    "coordinates": [
      {
        "$numberDouble": "-122.431272"
      },
      {
        "$numberDouble": "37.778008"
      }
    ]
  }
}
```

Sample Json – Restaurants



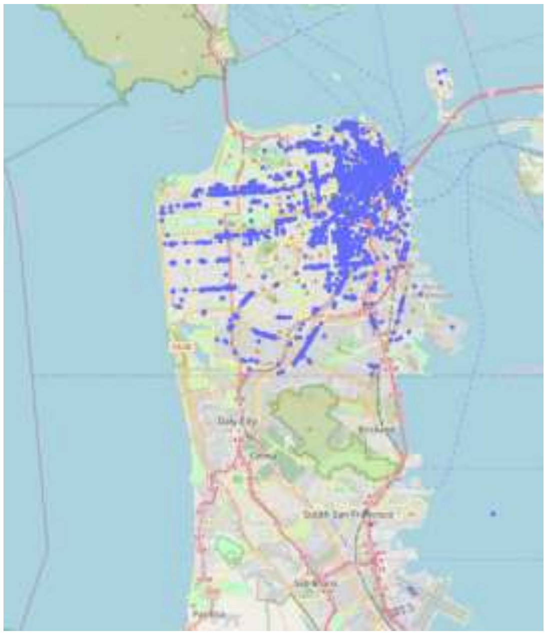
```
  _id: ObjectId('64052aec82dd2f2fa247d340')
  restaurant_name: "1. Mersea Restaurant & Bar"
  restaurant_url: "https://www.tripadvisor.com/Restaurant_Review-g60713-d13497261-Reviews…"
  restaurant_rating: "5.0 of 5 bubbles"
  restaurant_reviews_count: 543
  cost: "$$ - $$$"
  cuisine: "Vegetarian Friendly"
  rank: 1
  rest_address: "699 Avenue of the Palms Treasure Island, San Francisco, CA 94130"
  rest_phone: "+1 415-999-9836"
  rest_website: "http://www.mersea.restaurant/about-2/"
▾ geo_location: Object
    type: "Point"
  ▾ coordinates: Array
      0: -122.370818
      1: 37.824335
```

Sample Json Tourist Destination:

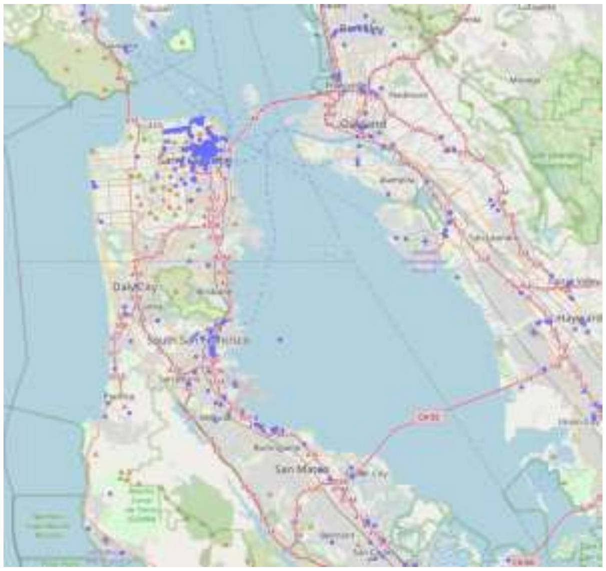


```
"15": {
  "Name": "Walt Disney Family Museum",
  "URL": "https://www.tripadvisor.com/Attraction_Review-g60713-d1556974-Reviews-Walt_Disney_Family_Museum-
San_Francisco_California.html",
  "TA Rating": "4.5 of 5 bubbles",
  "Num Reviews": "2,808",
  "Image": "https://dynamic-media-cdn.tripadvisor.com/media/photo-o/0c/64/fb/aa/photo0jpg.jpg?w=500&h=-1&s=1",
  "Description": "Speciality Museums",
  "Admission Cost": "from $25.00",
  "Address": "104 Montgomery St The Presidio, San Francisco, CA 94129-1718",
  "Geolocation": {
    "Longitude": -105.048897,
    "Latitude": 40.572029
  },
  "Write Review": "/UserReview-g60713-d1556974-Walt_Disney_Family_Museum-San_Francisco_California.html",
  "Review": {
    "1": {
      "Title": "What a surprise",
      "Review": "This little gemstone of a museum was quite a surprise treat. We have a membership at a museum local to us
that allows us to visit certain out of town museums for free. I found this one on the list and having been to San Fran a bunch
of times and seen all the attractions, said \"Why not.\"  \n\nThe first room contains Walt Disney's many many awards.  It's
more interesting than it sounds but don't get hung up here because the good stuff is further on.  Learn the story of how
Disney got started, and the work he did on the way to becoming a household name, and just how Micky Mouse got his start.  We
only allowed about an hour to see the museum and we were sorry as we needed at least twice that.  We will definitely go
back.Read more"
    },
    "2": {
      "Title": "Good for the Disney Enthusiast",
      "Review": "We found the museum informative and learned and saw things we hadn't seen before.  From original renderings,
cameras, and even legal papers that the Disney family worked with.  The area around the museum is also very nice - great views
of the GGB.  We came around opening and were able to find street parking.  Staff was friendly, and was limiting number of
people at a time so be prepared to wait unless you have prior reservations.  Read more"
    },
```

Sample Restaurant View:



Sample Hotel View:


Berkeley
Oakland
Daly City
Brisbane
South San Francisco
Pacifica
San Mateo
Hayward